ETH zürich

Molecular Descriptors

Theory and tips for real-world applications

Francesca Grisoni

University of Milano-Bicocca, Dept. of Earth and Environmental Sciences, Milan, Italy

ETH Zurich, Dept. of Chemistry and Applied Biosciences, Zurich, Switzerland

francesca.grisoni@unimib.it

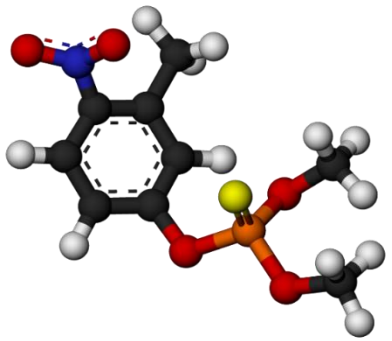
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

MODLAB

Presentation Outline

- Introduction
- Molecular representation and Molecular description
- Classical *vs* Fingerprint approach
- Tips and tricks

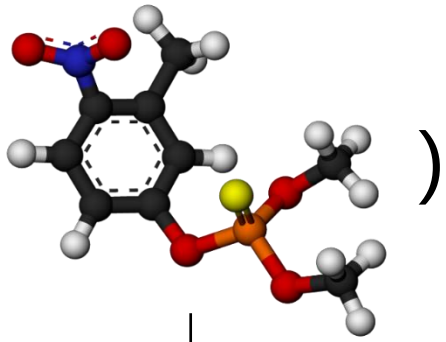
“It is obvious that there must exist a relation between the chemical constitution and the physiological action of a substance […], but as yet scarcely any attempts have been made to discover what this relation is. […] it might be supposed that a careful examination and comparison of known facts would lead to the discovery of some empirical law by means of which we **could deduce the action from the chemical constitution**.”

$P = f($  $)$

- Anesthetic potency *vs* oil/water partition coefficient (Meyer, 1899)
- Narcosis *vs* chain length (Overton, 1901)
- Narcosis *vs* surface tension (Traube, 1904)

Brown, A. C., & Fraser, T. R. (**1868**). Journal of anatomy and physiology, 2(2), 224.

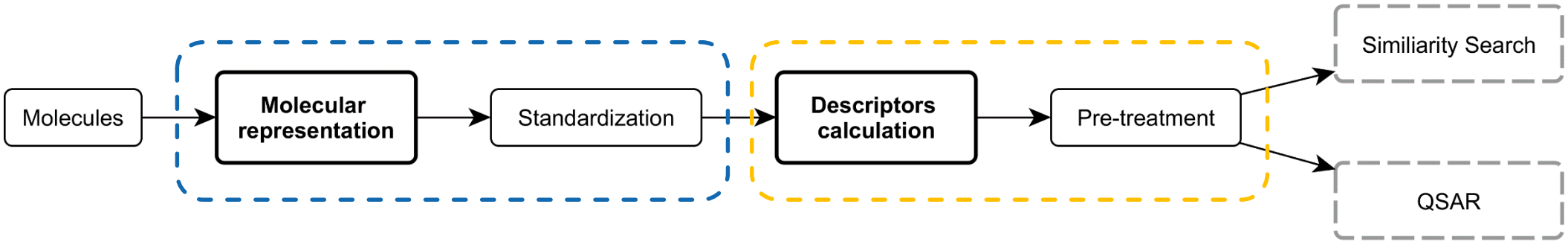
“… the final result of **a logical and mathematical procedure** that transforms **chemical information** of a molecule, such as structural features**, into useful numbers or the result of standardized experiments**.”

$$P = f\left(\text{  } \right)$$

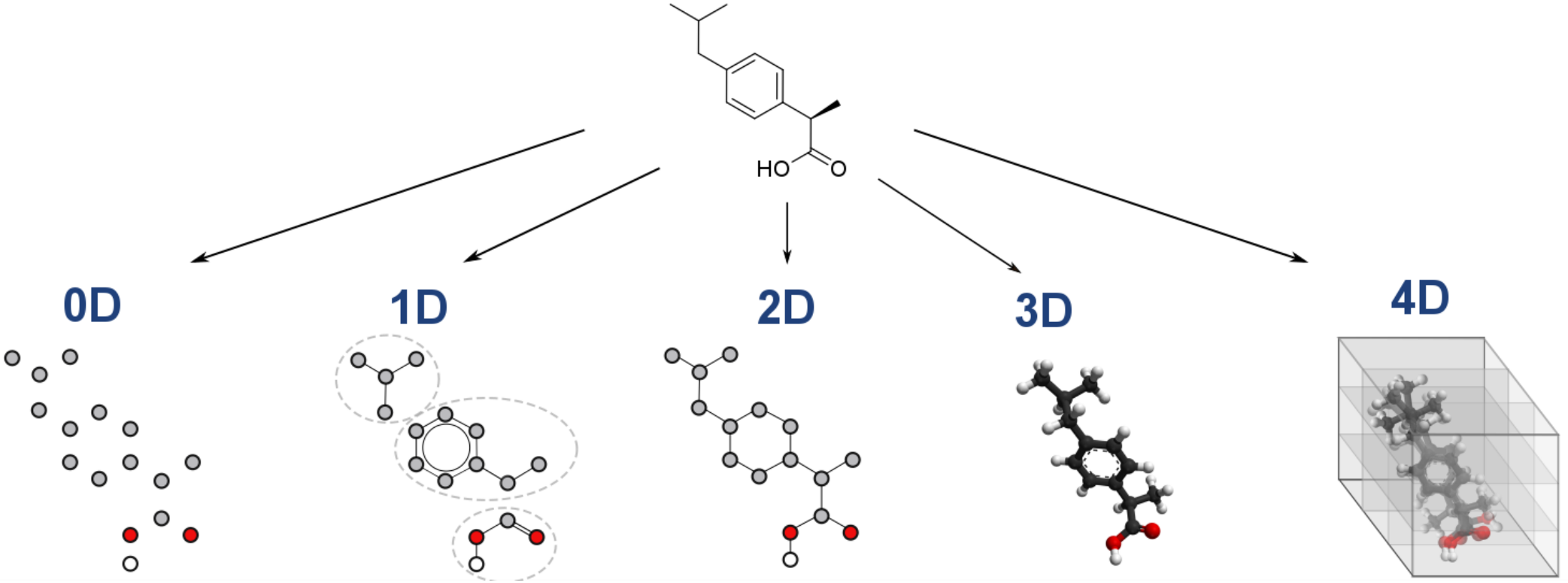
$$P = f\left(0.1, 1, 0, 3, 3.5, 100, 2, \ldots \right) + \varepsilon$$

Todeschini, R. & Consonni, V. (**2000**). *Handbook of molecular descriptors.* Wiley-VCH.

"… the final result of **a logical and mathematical procedure** that transforms **chemical information** of a molecule, such as structural features**, into useful numbers or the result of standardized experiments**."



Todeschini, R. & Consonni, V. (**2000**). *Handbook of molecular descriptors.* Wiley-VCH.



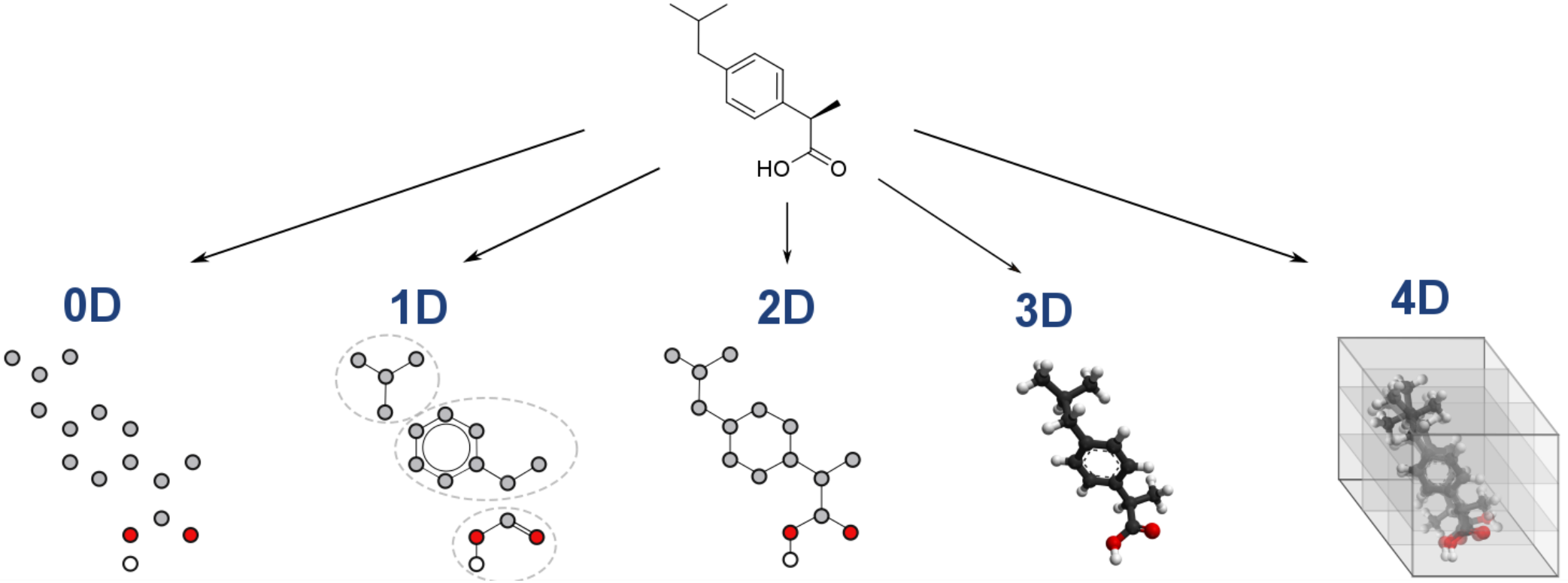
- Atom counts
- Molecular weight
- Atomic properties

- Fragment counts
- Fragment presence

- Topo-structural
- Topo-chemical

- Geometrical
- Atomic coordinates

- Grid-based
- Ensemble-based



- Atom counts
- Molecular weight
- Atomic properties

- Fragment counts
- Fragment presence

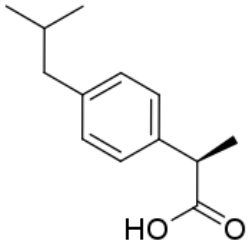
- Topo-structural
- Topo-chemical

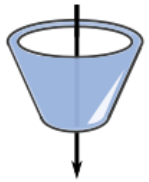
- Geometrical
- Atomic coordinates

- Grid-based
- Ensemble-based

"Make things as simple as possible, but not simpler."

Classical MDs

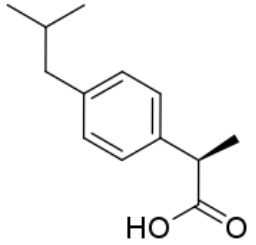




1

VS

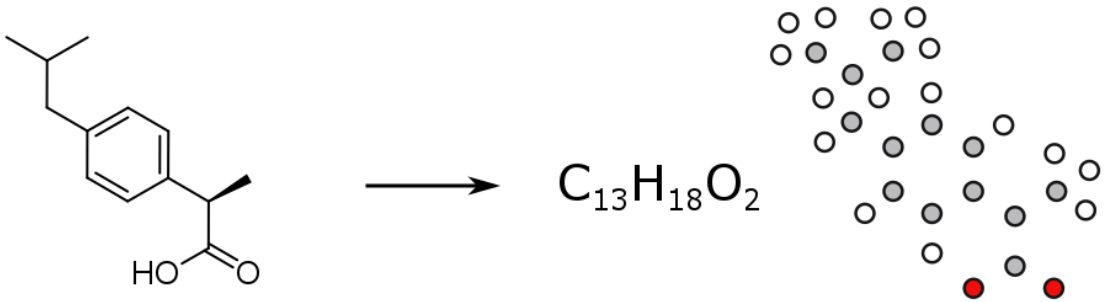
Fingerprints



1 0 0 0 1 0 ..... 0 1 1 0 0

n

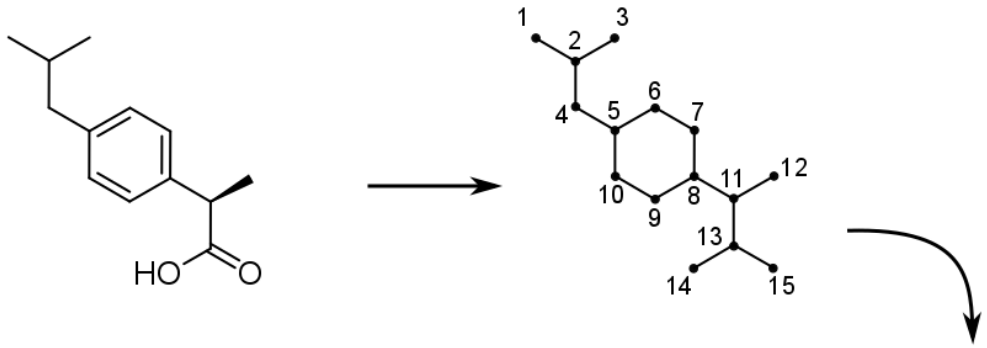
Molecular Weight



C 12.01 x 13 +
H 1.01 x 18 +
O 16.00 x 2 =

MW
206.31

Matrix-based descriptors

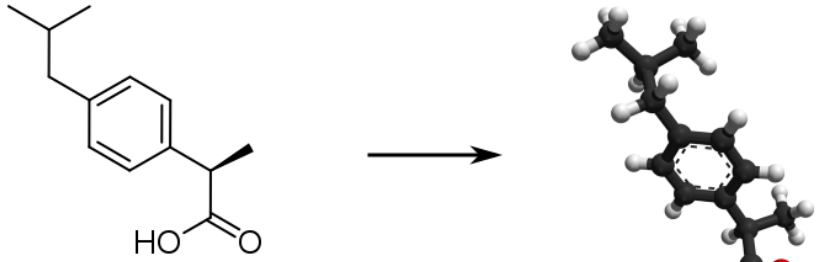


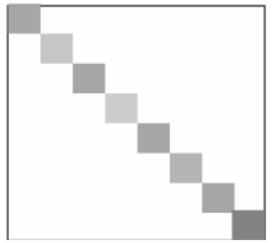
distance/detour matrix (D/D_t)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	O	O
C1	0	1	1	1	1	0.5	0.7	1	0.7	0.5	1	1	1	1	1
C2	1	0	1	1	1	0.4	0.7	1	0.7	0.4	1	1	1	1	1
C3	1	1	0	1	1	0.5	0.7	1	0.7	0.5	1	1	1	1	1
C4	1	1	1	0	1	0.3	0.6	1	0.6	0.3	1	1	1	1	1
C5	1	1	1	1	0	0.2	0.5	1	0.5	0.2	1	1	1	1	1
C6	0.5	0.4	0.5	0.3	0.2	0	0.2	0.5	1	0.5	0.6	0.7	0.7	0.7	0.7
C7	0.7	0.7	0.7	0.6	0.5	0.2	0	0.2	0.5	1	0.3	0.4	0.4	0.5	0.5
C8	1	1	1	1	1	0.5	0.2	0	0.2	0.5	1	1	1	1	1
C9	0.7	0.7	0.7	0.6	0.5	1	0.5	0.2	0	0.2	0.3	0.4	0.4	0.5	0.5
C10	0.5	0.4	0.5	0.3	0.2	0.5	1	0.5	0.2	0	0.6	0.7	0.7	0.7	0.7
C11	1	1	1	1	1	0.6	0.3	1	0.3	0.6	0	1	1	1	1
C12	1	1	1	1	1	0.7	0.4	1	0.4	0.7	1	0	1	1	1
C13	1	1	1	1	1	0.7	0.4	1	0.4	0.7	1	1	0	1	1
O	1	1	1	1	1	0.7	0.5	1	0.5	0.7	1	1	1	0	1
O	1	1	1	1	1	0.7	0.5	1	0.5	0.7	1	1	1	1	0

λ_{MAX} 11.21

GEometry, Topology, and Atom-Weights Assembly (GETAWAY)

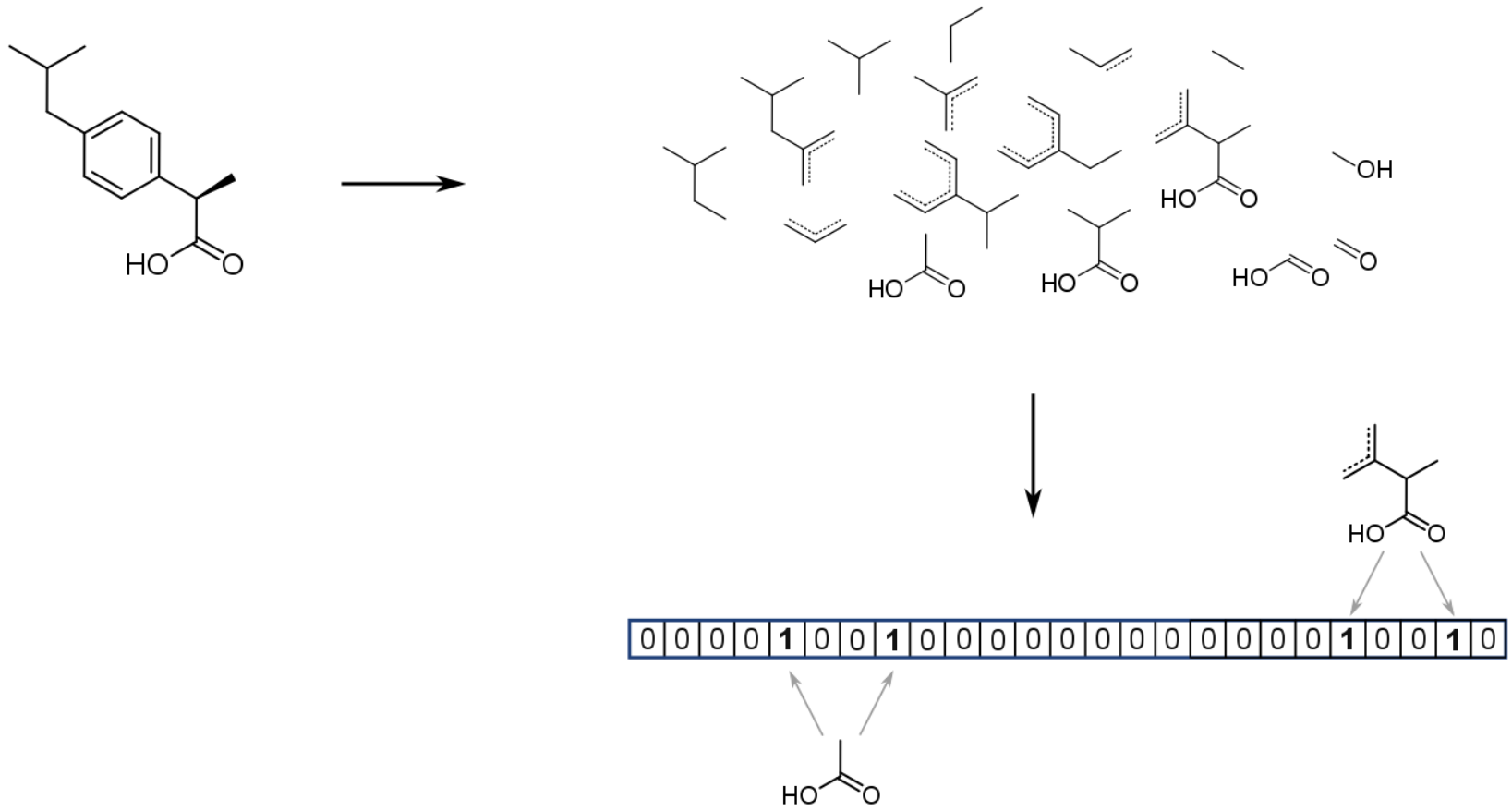




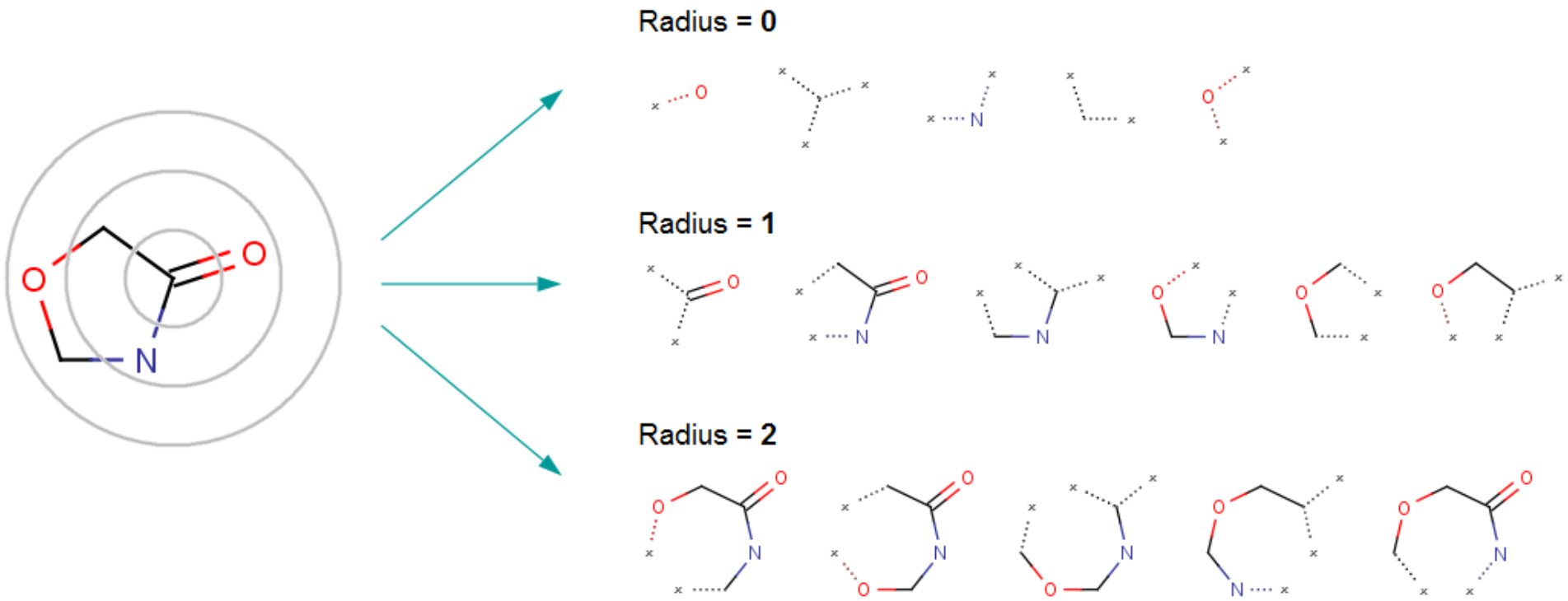
$$\mathbf{H} = \mathbf{M} \times (\mathbf{M}^{\mathbf{T}} \times \mathbf{M})^{-1} \cdot \mathbf{M}^{\mathbf{T}}$$

$$HIC = -\sum_{i=1}^{nAT} \frac{h_{ii}}{D} \cdot \log_2 \frac{h_{ii}}{D}$$

Binary Fingerprints

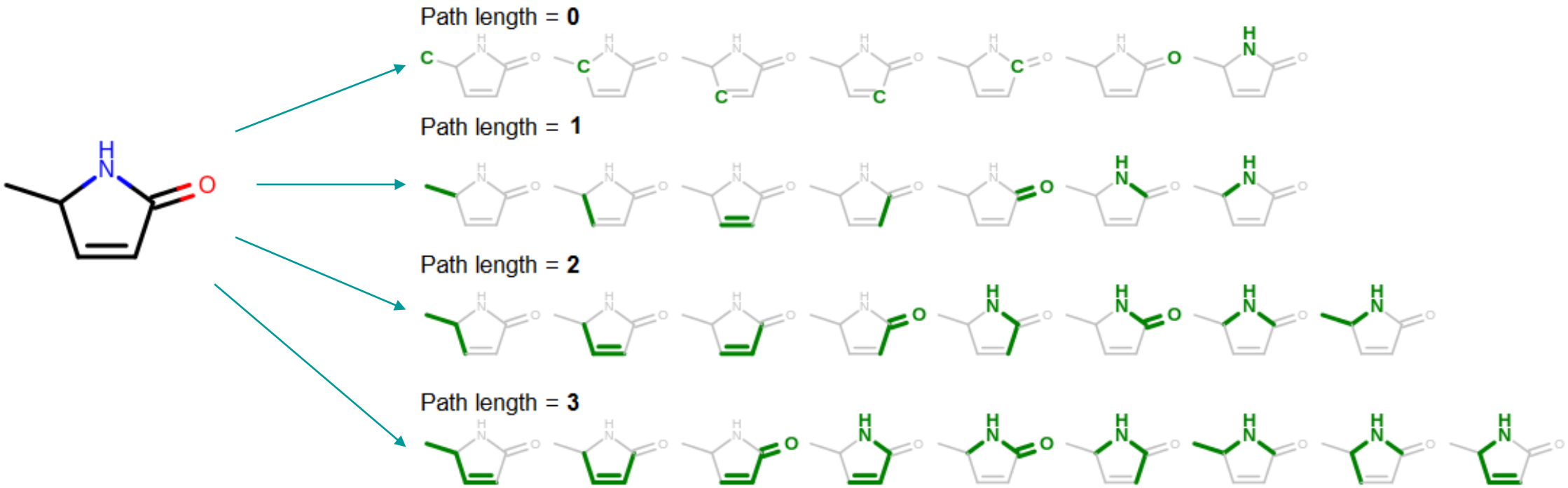


Extended Connectivity FP



https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP

Path FP

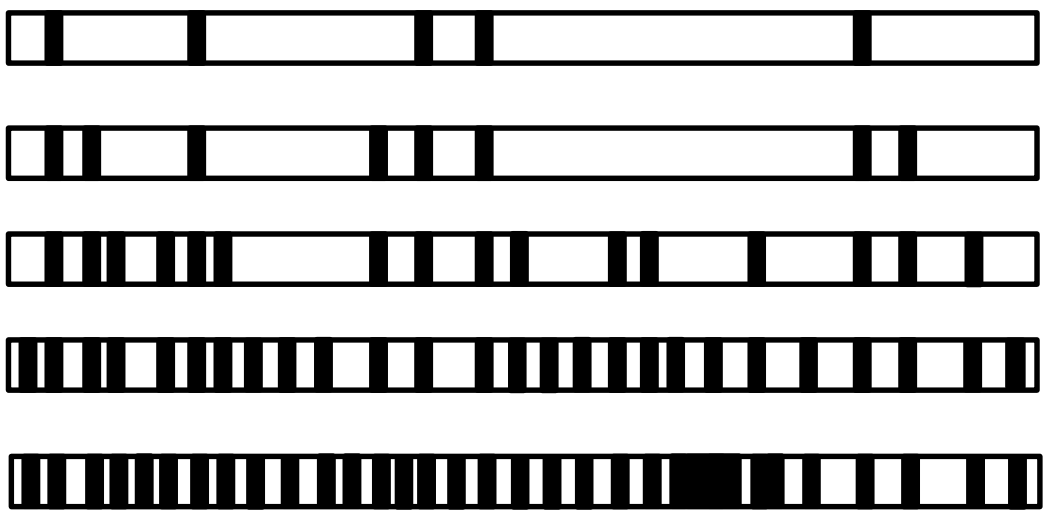


https://docs.eyesopen.com/toolkits/python/graphsimtk/fingerprint.html

FP settings

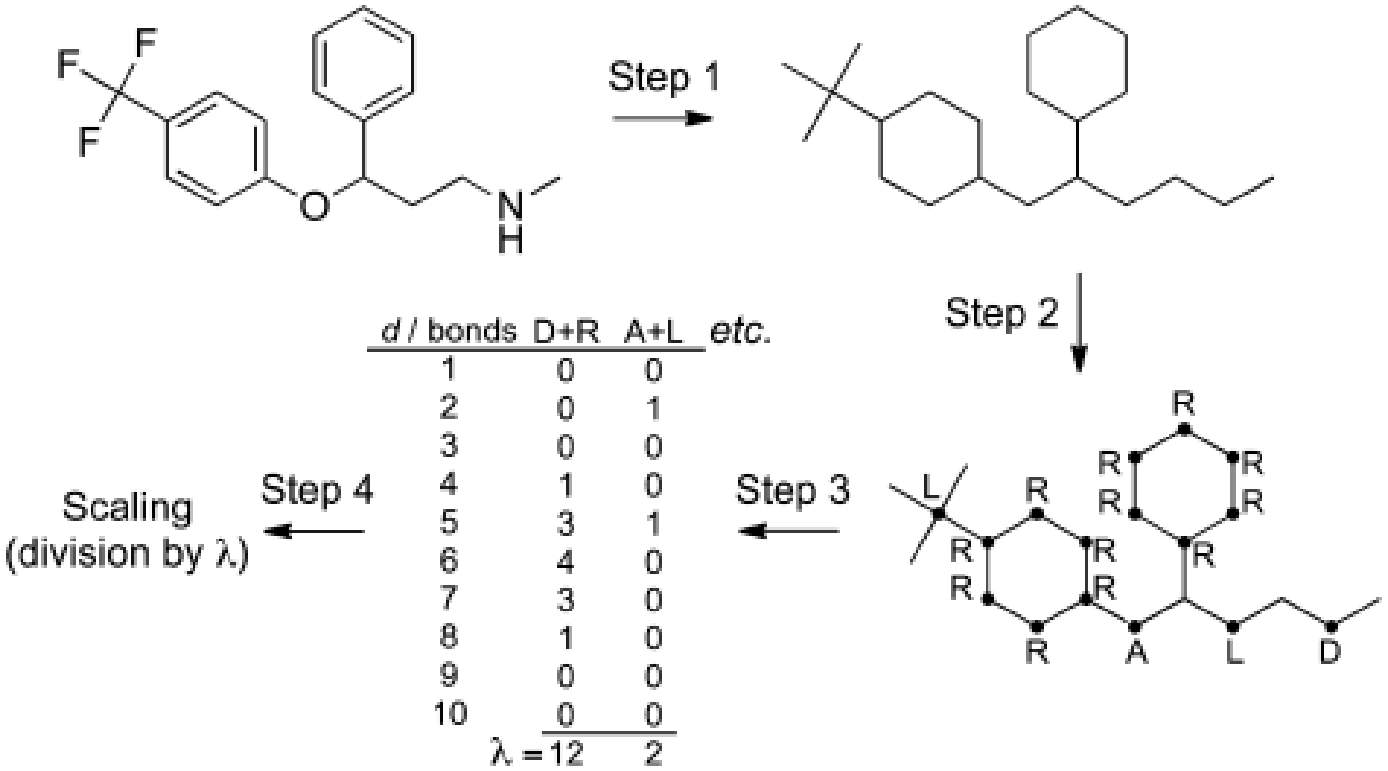
- Radius/path length
- Number of bits
- FP length

- Molecular Information
- Bit collision
- Darkness



Darkness (av. 40-50%, max 80%)

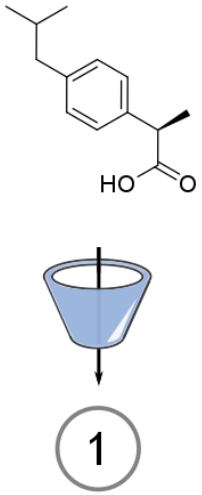
Chemically Advanced Template Search (CATS)



Reutlinger, M., Koch, C. P., Reker, D., Todoroff, N., Schneider, P., Rodrigues, T., & Schneider, G. (2013). *Mol. Inf.* 32(2), 133-138.

Which approach?

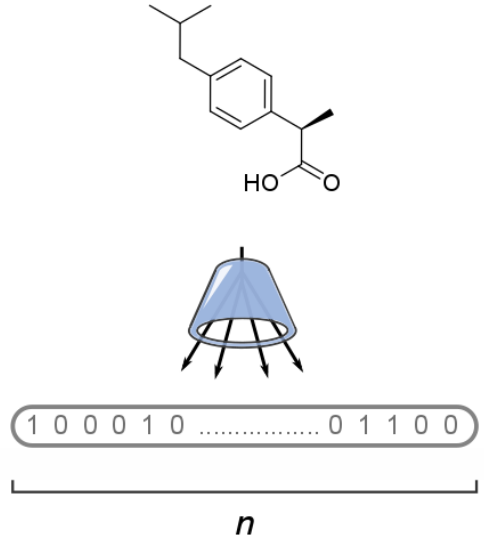
Classical MDs



- Amount of encoded information
- Interpretability
- Require pre-treatment

VS

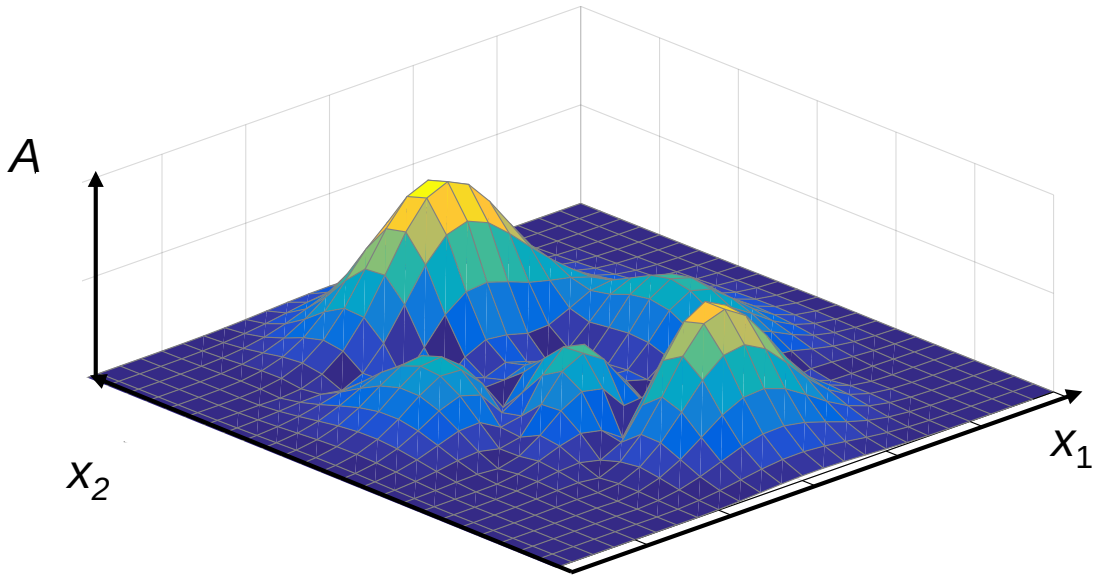
Fingerprints



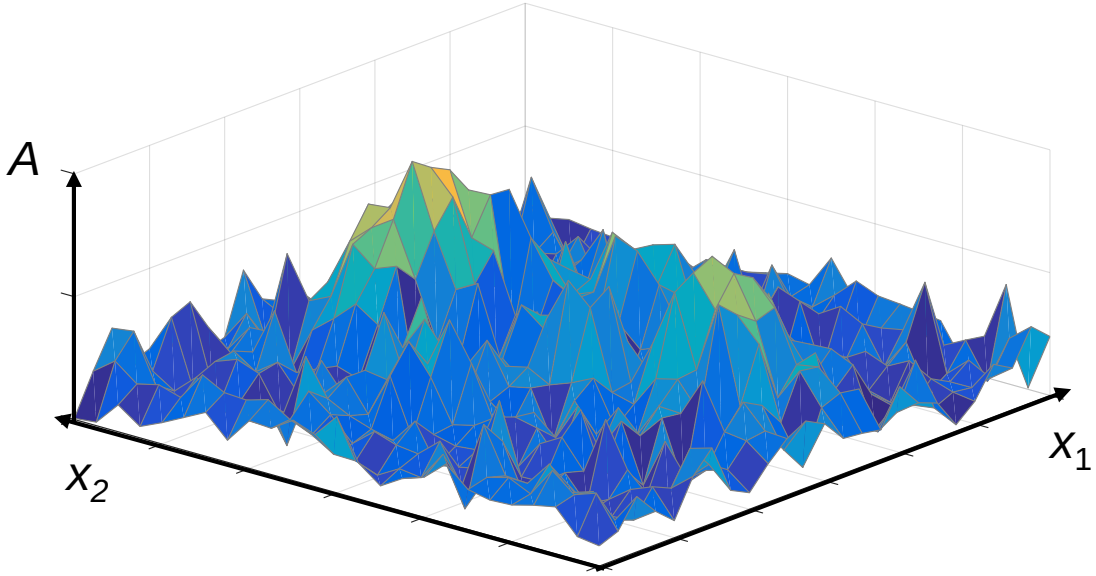
- Quick similarity calculations
- No need for pre-treatment
- Modelling approaches for binary data

Structure Activity Landscapes

gently rolling hills

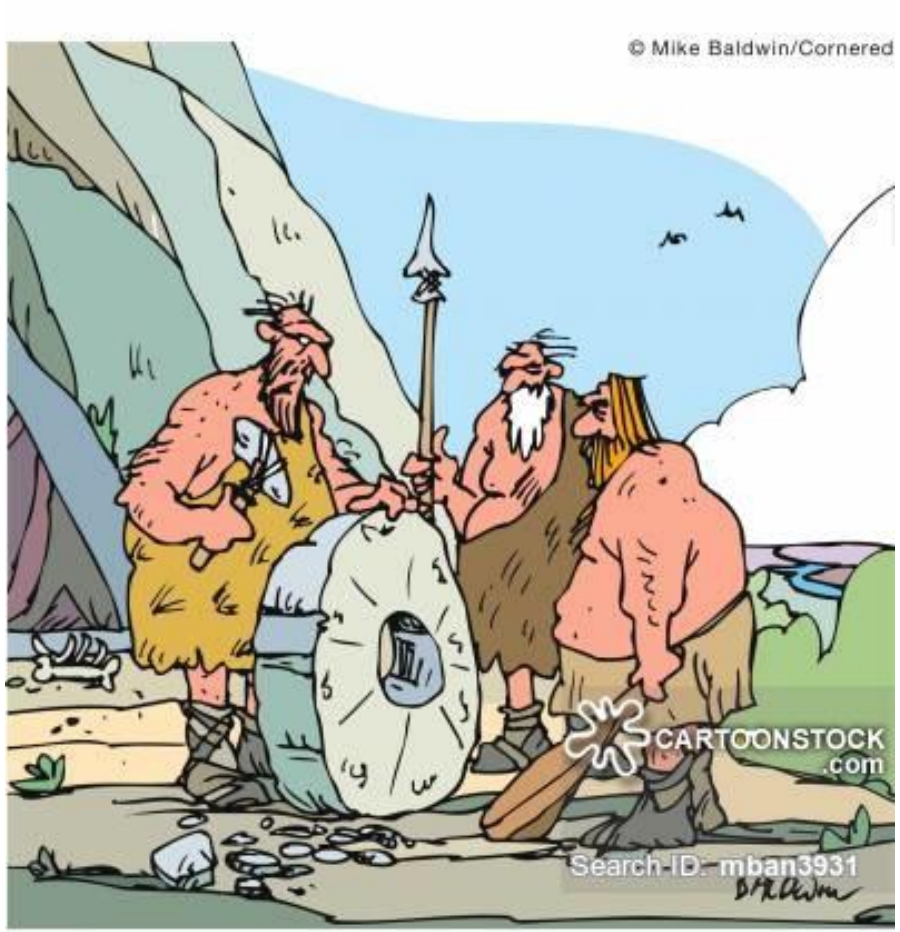


rugged landscapes



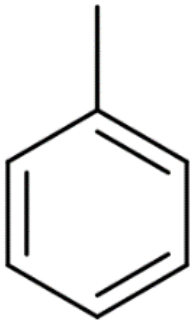
$$SALI_{i,j} = \frac{|A_i - A_j|}{1 - sim(i, j)}$$

Tips and tricks



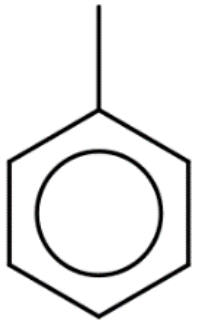
Tips and tricks

0. Attention to structure representation



SMILES: CC1=CC=CC=C1

nBM = 3

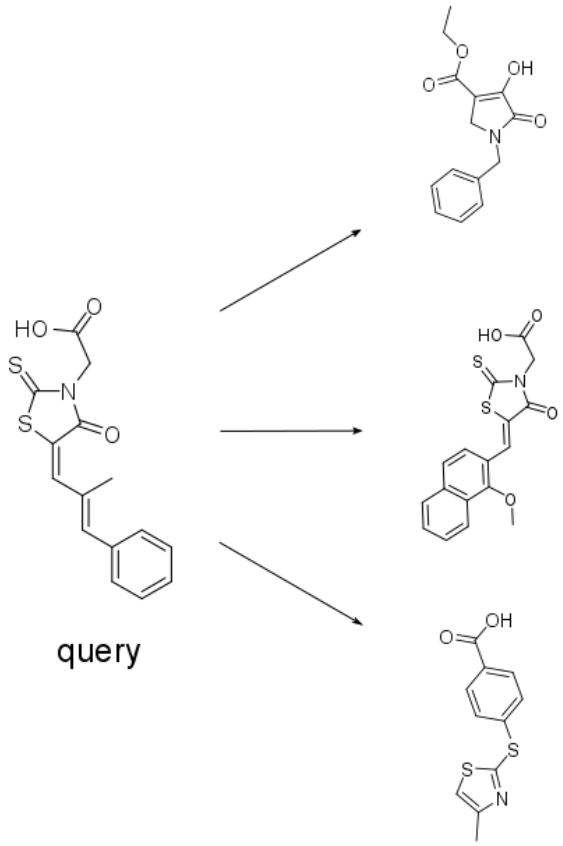


SMILES: Cc1ccccc1

nBM = 6

Tips and tricks

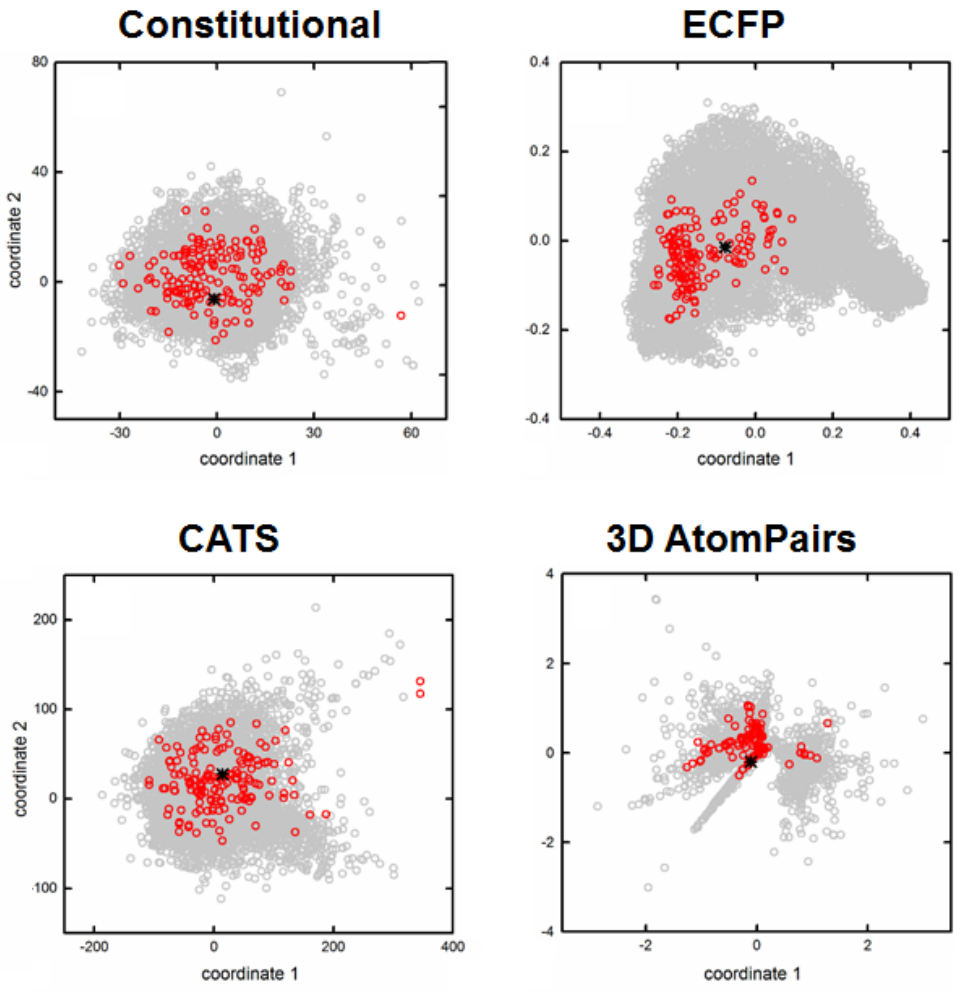
1. Know your purpose



Functional group counts

CATS

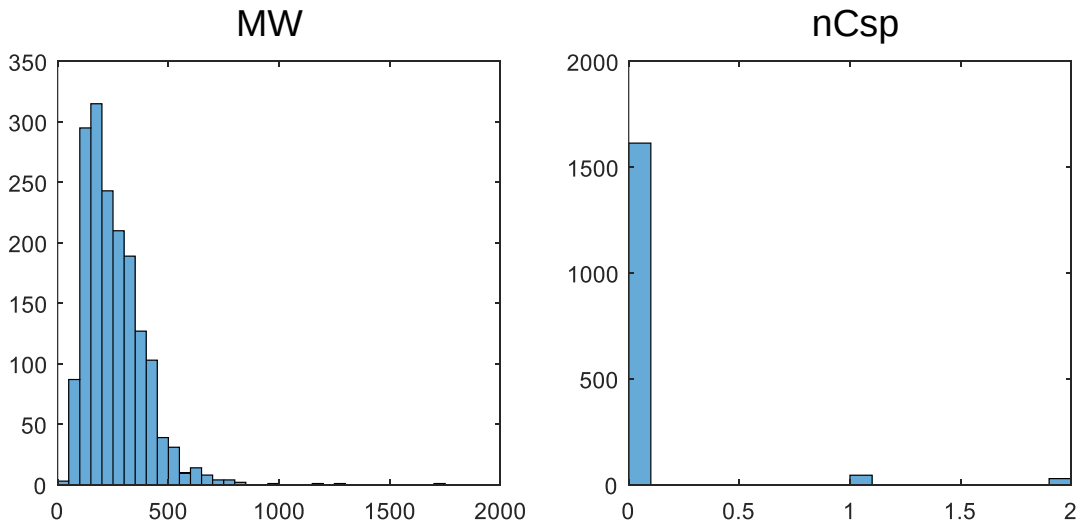
3D AtomPairs

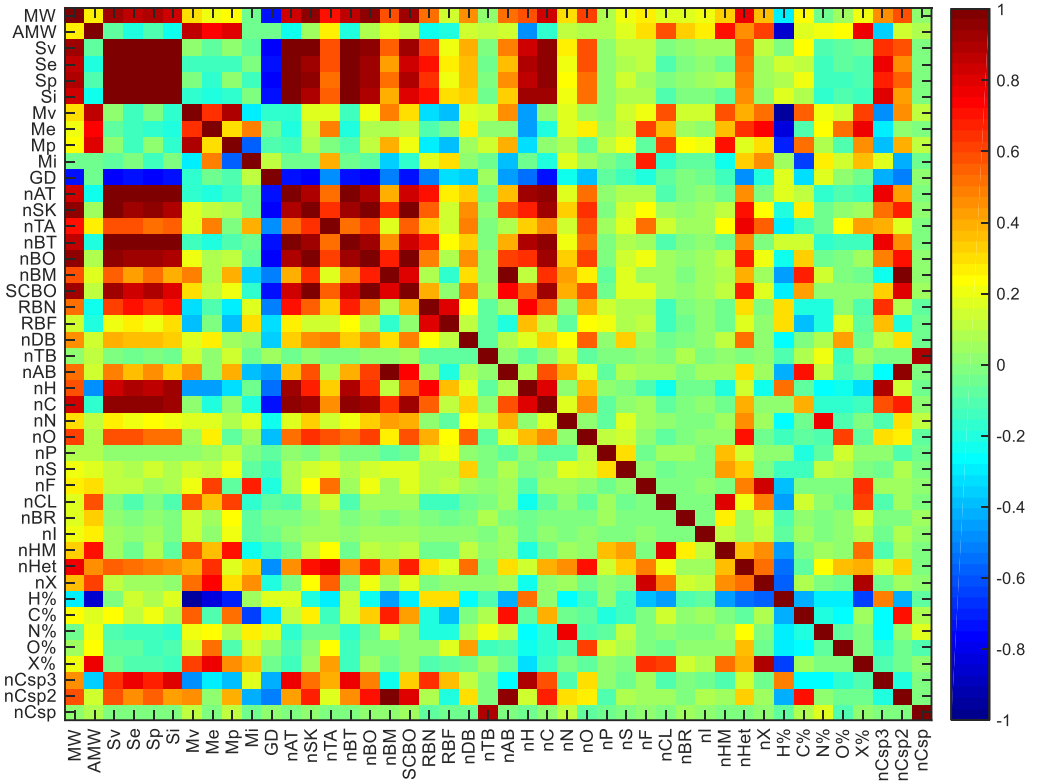


Grisoni, F., Consonni, V., Todeschini, R. (2017). Impact of molecular descriptors on computational models. In *Computational Chemogenomics*, Methods in Molecular Biology, Springer. (In press)

Tips and tricks

2. Reduce Dimensionality (if possible)

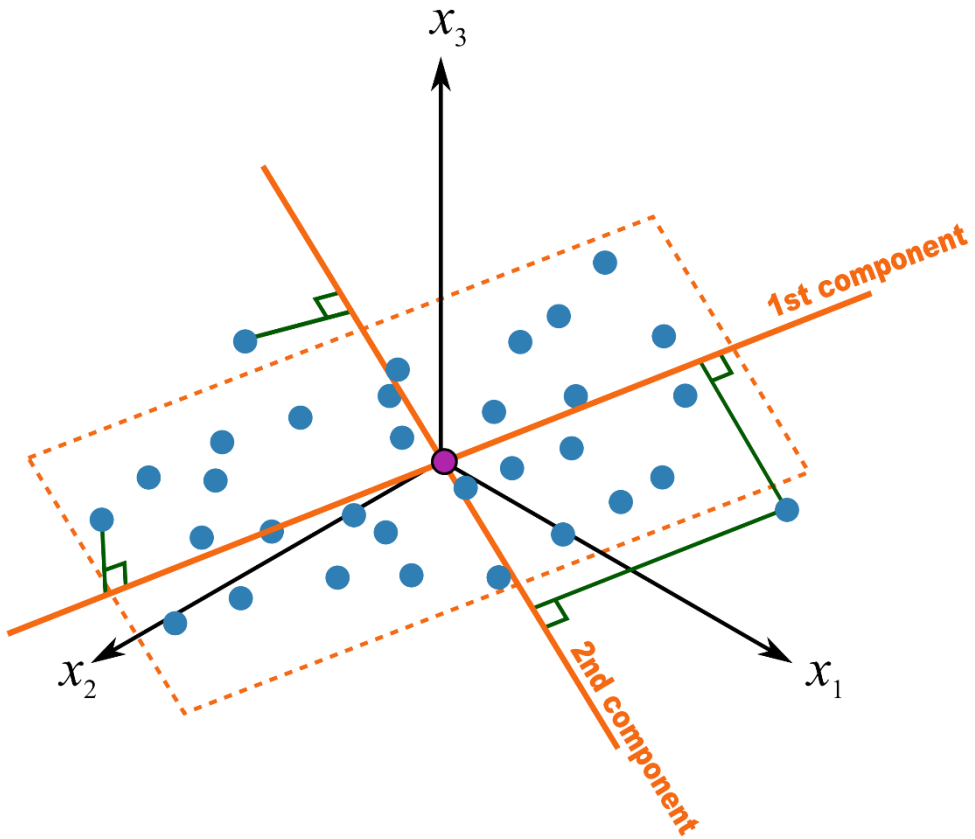




Tips and tricks

2. Reduce Dimensionality (if possible)

PCA = Principal Component Analysis



https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/geometric-explanation-of-pca

Tips and tricks

2. Reduce Dimensionality (if possible)

K-means clustering

- Assign variables randomly to a set of *k* clusters
- Compute cluster centroids
- Re-assign variables to the cluster with the closest centroid

Tips and tricks

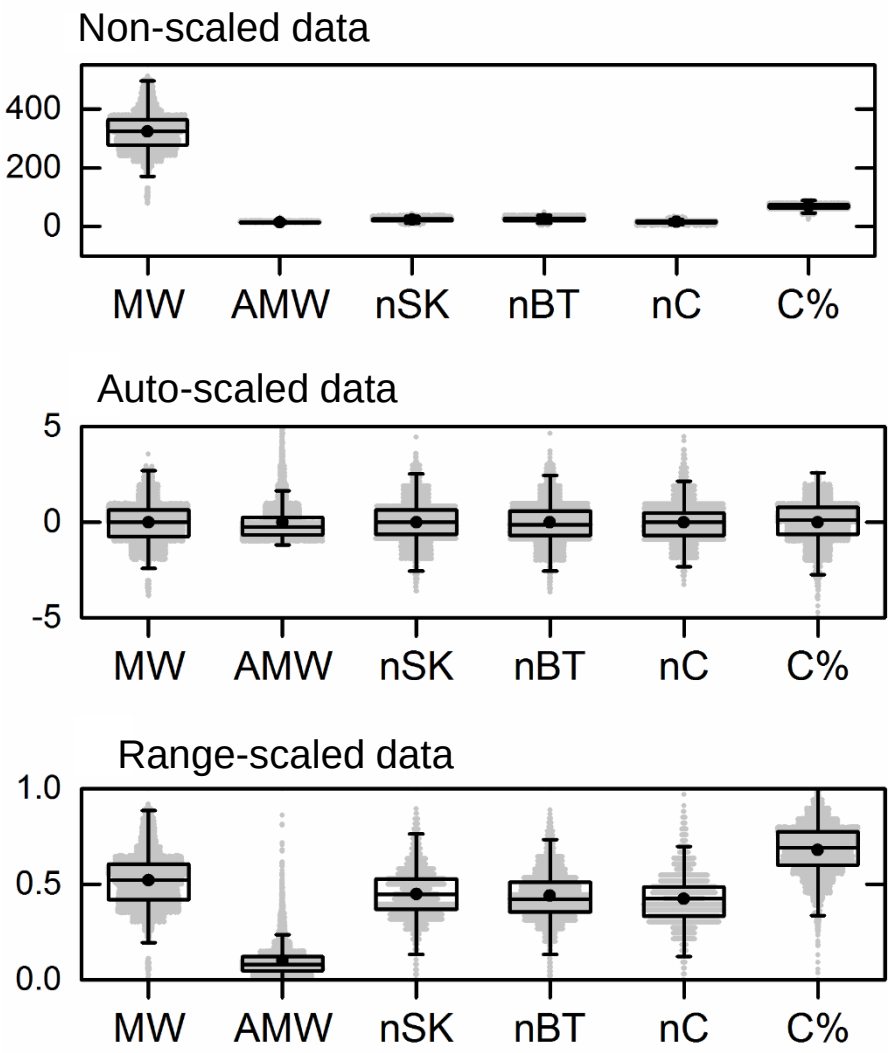
3. Mind the measuring unit

- **Auto-scaling** (Gaussian normalization)

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

- **Range-scaling** (minMax normalization)

$$x'_{ij} = \frac{x_{ij} - \min_j}{\mathrm{Max}_j - \min_j}$$

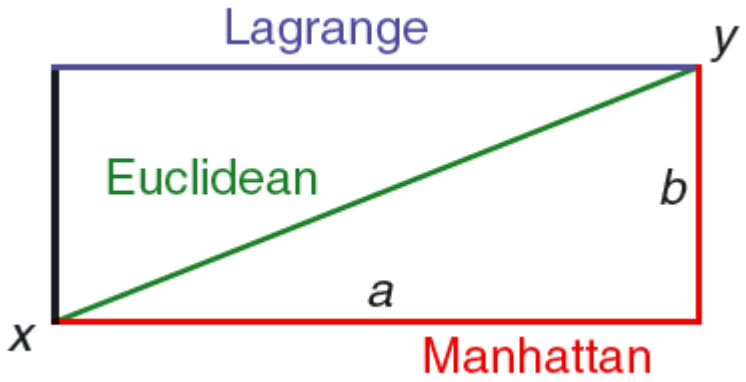


Grisoni, F., Consonni, V., Todeschini, R. (2017). Impact of molecular descriptors on computational models. In *Computational Chemogenomics*, Methods in Molecular Biology, Springer. (In press)

Tips and tricks

4. Consider other similarity measures

Distance	Definition
Euclidean	$D_{xy}^{\mathrm{EUC}} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Manhattan or city-block	$D_{xy}^{\mathrm{MAN}} = \sum_{j=1}^{p}\lvert x_j - y_j\rvert$
Lagrange	$D_{xy}^{\mathrm{LAG}} = \max_j\lvert x_j - y_j\rvert$
Minkowski	$D_{xy}^{\mathrm{MIN}} = \left[\sum_{j=1}^{p}\left\lvert x_j - y_j\right\rvert^q\right]^{1/q}$
Mahalanobis	$D_{xy}^{\mathrm{MAH}} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\cdot\mathbf{S}^{-1}\cdot(\mathbf{x}-\mathbf{y})}$



Todeschini, R., Ballabio, D., & Consonni, V. (2015). Distances and other dissimilarity measures in chemometrics. *Encyclopedia of analytical chemistry*.

Tips and tricks

4. Consider other similarity measures

Similarity coefficient	Definition
Sokal–Michener, Simple Matching	$S_{xy}^{SM} = \frac{a+d}{p}$
Rogers–Tanimoto	$S_{xy}^{RT} = \frac{a+d}{p+b+c}$
Jaccard–Tanimoto	$S_{xy}^{RT} = \frac{a}{a+b+c}$
Gleason–Dice–Sorensen	$S_{xy}^{GLE} = \frac{2a}{2a+b+c}$
Russell–Rao	$S_{xy}^{RR} = \frac{a}{p}$
Forbes	$S_{xy}^{FOR} = \frac{pa}{(a+b)(a+c)}$
Simpson	$S_{xy}^{SIM} = \frac{a}{\min\{(a+b),(a+c)\}}$
Braun-Blanquet	$S_{xy}^{BB} = \frac{a}{\max\{(a+b),(a+c)\}}$
Driver–Kroeber–Ochiai cosine	$S_{xy}^{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$
Baroni-Urbani–Buser	$S_{xy}^{BU1} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
Kulczynski	$S_{xy}^{KUL} = \frac{1}{2} \cdot \left[\frac{a}{a+b} + \frac{a}{a+c}\right]$

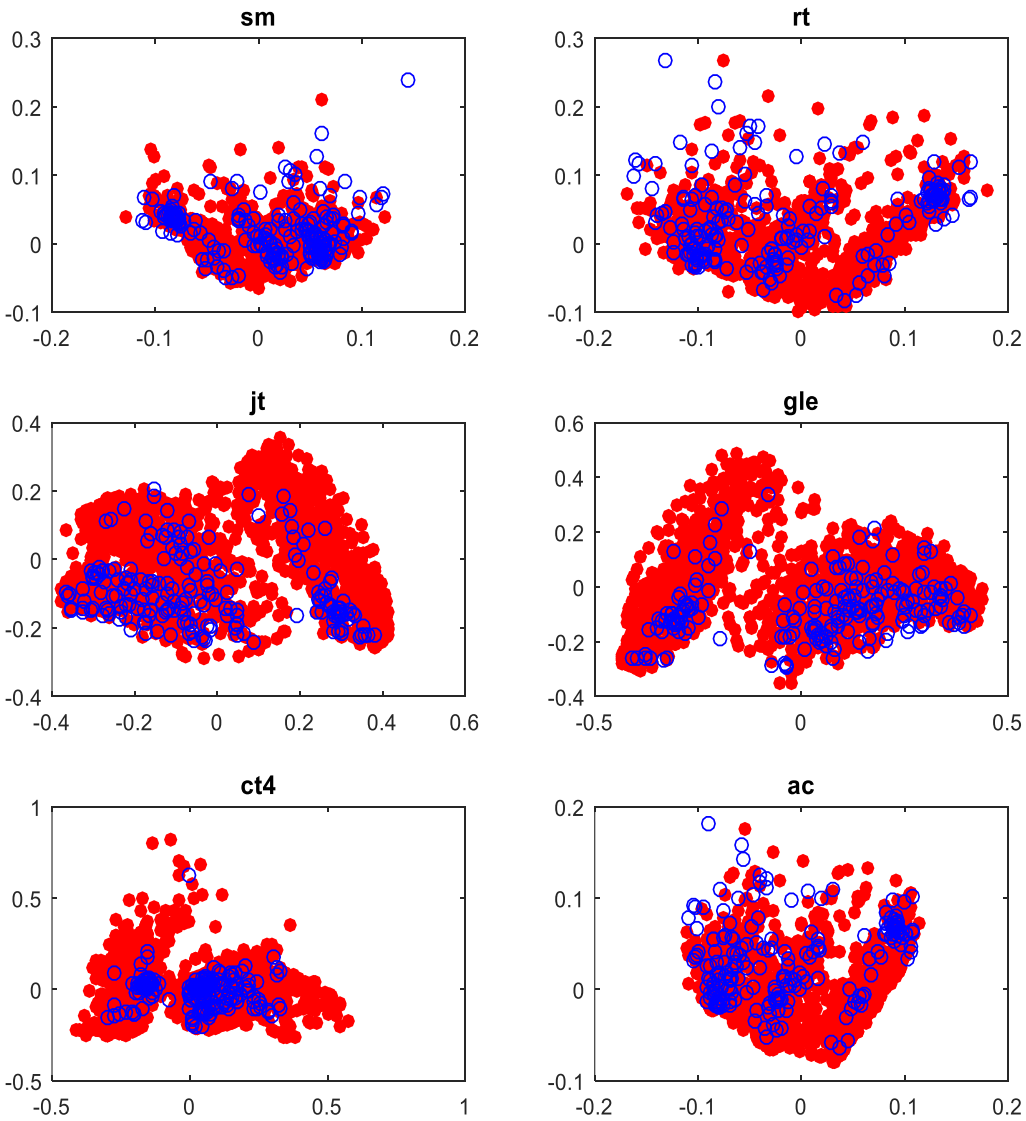
	1	0
1	a	b
0	c	d

Todeschini, R., Ballabio, D., & Consonni, V. (2015). Distances and other dissimilarity measures in chemometrics. *Encyclopedia of analytical chemistry*.

Tips and tricks

4. Consider other similarity measures

Similarity coefficient	Definition
Sokal–Michener, Simple Matching	$S_{xy}^{SM} = \frac{a+d}{p}$
Rogers–Tanimoto	$S_{xy}^{RT} = \frac{a+d}{p+b+c}$
Jaccard–Tanimoto	$S_{xy}^{RT} = \frac{a}{a+b+c}$
Gleason–Dice–Sorensen	$S_{xy}^{GLE} = \frac{2a}{2a+b+c}$
Russell–Rao	$S_{xy}^{RR} = \frac{a}{p}$
Forbes	$S_{xy}^{FOR} = \frac{pa}{(a+b)(a+c)}$
Simpson	$S_{xy}^{SIM} = \frac{a}{\min\{(a+b),(a+c)\}}$
Braun-Blanquet	$S_{xy}^{BB} = \frac{a}{\max\{(a+b),(a+c)\}}$
Driver–Kroeber–Ochiai cosine	$S_{xy}^{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$
Baroni-Urbani–Buser	$S_{xy}^{BU1} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
Kulczynski	$S_{xy}^{KUL} = \frac{1}{2} \cdot \left[\frac{a}{a+b} + \frac{a}{a+c}\right]$



Todeschini, R., Ballabio, D., & Consonni, V. (2015). Distances and other dissimilarity measures in chemometrics. *Encyclopedia of analytical chemistry*.

Summary

- Descriptors are numbers that capture particular molecular features
- The best descriptors set depends on the problem
- Different types of descriptors require different type of pre-treatment
- Molecular similarity is not an absolute concept

Additional Reading

- **Molecular descriptor theory**
  - Mauri, A., Consonni, V., Todeschini, R. (2016). Molecular descriptors. In *Handbook of Computational Chemistry*, Springer.

- **Tutorial on descriptors processing and use**
  - Grisoni, F., Consonni, V., Todeschini, R. (2017). Impact of molecular descriptors on computational models. In *Computational Chemogenomics*, Methods in Molecular Biology, Springer. (*In press*)

- **Automated data pre-processing**
  - Mansouri, K., Grulke, C.M., Richard A.M., *et al.* (2016). An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling, SAR and QSAR in Environmental Research. 27, 911–937.

Software (some examples)

Software	No. descr.	Description	Free
ADMEWORKS ModelBuilder	≈ 400	Physicochemical, topological, geometrical, and electronic properties derived from the molecular structure	
BlueDesc	174	Descriptors from JOELib2 and CDK sources, works only with 3D structures.	
CODESSA	≈ 1,500	Constitutional, topological, geometrical, charge-related, quantum-chemical and thermodynamic descriptors.	
Dragon	> 5,200	Benchmark software for calculating 0- to 3D descriptors and binary fingerprints.	
E-Dragon	> 3,000	Free, electronic remote version of DRAGON.	yes
MOE - Molecular Operating Environment	≈ 300	Topological indices, structural keys, E-state indices, physical properties.	
PaDel	> 1,875	Open source. Based on CDK with additional 2D and 3D descriptors.	yes
ISIDA Fragmentor	/	Molecular fragments from a Structure-Data File (SDF).	